CLASSIFYING FORMAL AND INFORMAL WORKERS IN COLOMBIA
Machine Learning - Final Project Report

Carmen Garro
C.garro@hertie-school.org

María José Lee
M.lee@hertie-school.org

Gresa Smolica
G.smolica@hertie-school.org

Abstract

This paper[1] *documents the classification of formal and informal workers in the Colombian Labour Market using the country's Large Integrated Household Survey (GEIH by its initials in Spanish)[3]*[2]*. While historically Colombia has maintained a moderately high rate of informality, the pandemic and its subsequent shocks had a negative effect on society as a whole, making unemployment rates peak, and driving more people toward informality. All of this had a significant impact on the labour conditions, social contributions, and in general, the well-being of workers, both in the formal and informal market.*

The aim is to bring an innovative and useful policy-making tool to address the informal economy, by illustrating how machine learning classification techniques can significantly benefit the public policy area, not only in policy-design, but also in policy evaluation and implementation[8]. Among the main findings shown by our models is the strong relation between the variables associated to particular Working Characteristics of respondents and the likelihood of being classified as an informal or formal worker.

1. *Introduction

The Informal Economy can be understood as the goods and services that are purchased or contracted outside of formal jurisdiction in a country or economy. Because of its hidden nature, governments tend to have a hard time dealing with it, as well as the consequences that emerge from it. Amongst the many implications of this phenomenon, one of the main problems is the Fiscal Aspect, as that which is not recorded cannot be taxed, in addition to the added difficulty of ensuring a safe workspace and appropriate benefits for those involved.

Due to the overall negative effects that informality can have on an economy, having a clear understanding of how big the informal sector is, tends to be a priority for governments dealing with said challenge, in addition to understanding the factors that influence its overall size. However, considering the complexity of this task, a strong political will, combined with a thoroughly thought-out estimation process is crucial to carry this out effectively.

Amongst the countries with a higher percentage of informal workers within their economy, we find Colombia - a country that historically has had an informal workforce share of around half its working force, but for which, after the COVID-19 Pandemic, this has steadied closer to 60%. This phenomenon is usually attributed to a combination of two factors, (i) a difficulty for the formal labour market to adequately absorb willing workers into it, and (ii) a reluctance on the informal workers side to be on the government's radar in the first place.

Colombia also stands out amongst its Latin American peers by virtue of its commitment to providing open and public data, with the intention of allowing concerned citizens, researchers or other interested parties to take a closer look at the inner workings of Government, or alternatively, databases concerning the current state of infrastructure, the environment, the population, etc., amongst which we find the GEIH. Throughout our research paper, precisely such data is used to pick up on the manifestation of the informal economy, with the intention of it being able to be used towards generating better policies.

The reviewed literature acknowledges the difficulty of estimating the size of the informal economy for different countries, mentioning among the main hurdles, the lack or non-consistency of statistics that routinely cover households. Our approach, however, addresses this challenge directly, by exploiting the fact that this survey has been run on a consistent basis for the past 13 years, accumulating a significant amount of data, and thus, creating the opportunity for a representative estimation which will further

[1] The link to open the GitHub Repository is https://github.com/GresaSm/ML-Project-Year-I.git

[2] Gran Encuesta Integrada de Hogares

enrich the knowledge and the information on Informal Economy, serving then directly as a prediction tool, and indirectly, as a potential tool for better policy design.

Our project aims to build upon the existing research carried out by the International Labour Organisation, International Monetary Fund and the World Bank when tackling the informal economy. As an added value to such literature, our work contributes to a more narrow approach of tackling a public policy challenge with the support of machine learning models.

2. *Related Work

Until recent years, studies and research on informality have been conducted mostly from a qualitative perspective. Focusing mostly in developing countries, the Informal Economy has been seen as a phenomenon related to geographical characteristics (rural/urban) or by the state of development of the country where it manifests (developing/poor countries). Some estimate that by 2021, said sector constituted more than the 70 percent of total employment in these countries, and roughly one-third of output. According to the World Bank's report "The long shadow of informality", informal businesses constitute 72 percent of firms in the services sector, its workers are predominantly women, and they are usually young and low-skilled[9].

However, an increased interest in this topic shed further light on the complexity of the issue and the reality that came with it. The existence of undeclared work was evidenced in developed countries, together with the fact that this harmful practice was going to stay around for longer in the economies of the world[4]. Alogogianni et.al, argue that employing advanced machine learning techniques in the data produced by the labour inspectorates and the e-Government public services, could make up for the difficulties when it comes to institutions lacking human and financial resources to address the issue, but also in providing significant assistance in undeclared work prediction and understanding its features[4].

Similar to our algorithm model, the authors argue that classification algorithms may learn from datasets containing past labour inspection findings and produce classifiers that effectively predict labour law violations, as well as provide understandable explanations for these predictions However the challenge noticed throughout the reviewed literature lies in the hidden nature of the informal economy resulting in it being underrepresented in datasets (such as the labour institutions surveys, household surveys, etc.), a very important component that hinders the machine learning process. However, going back to the indicators set by several international organisations such as ILO, World Bank, etc., several features of the informal economy have been identified and used when predicting within this area. In his research paper, 'Informal Work in Numbers', Hummel employs a machine learning algorithm in the data derived from the Latin American Public Opinion Project. The used regression model suggested that informal workers are more likely to organise in low-capacity countries[5].

On a policy level, drawing back to Colombia, the arrival of COVID-19 disrupted about a quarter of employment. From a local perspective, or from the workers side, informality was both praised and criticised. While it helped the economy to bounce back from the pandemic devastation, in the labour conditions aspect, workers experienced and are experiencing lack of employment protection. Adjusting to informality projects longer-term costs, therefore it is crucial for such "flexibility"/adjustment not to de-prioritize designing policies that create a more formal economy in the medium-term[7]. Machine-learning tools seem to bring the needed innovation in learning from the process and helping to address the issue in the medium-long term.

3. Proposed Method

The applied classification model falls under Supervised Learning Methods. As such throughout the work the following models have been employed for prediction: the Logistic Regression, Decision Tree, and the XGBoost[6]. Following our experimenting phase where we tested these three approaches, the results showed us that the Decision Tree model is the one that fits the best to our case, despite the fact that all three models have shown good performance when trying to predict unseen data.

Throughout the work, the following steps were taken:

- Initially the team ran models in the dataset of 2019 while drawing a sample consisting of the first quarter of the year
- Further the model was trained to be applied in the annual dataset of 2019, which naturally consisted of more observations;
- Finally, the trained model was used/tested to predict the classification of informal/formal worker in the GEIH survey dataset for the year 2021.

Prior to running our baseline model, we analysed the class distribution to see whether data imputation or another data balancing technique was needed. As seen in *Figure 1* the proportion of the classes was sufficiently balanced, which allowed to continue running the models.

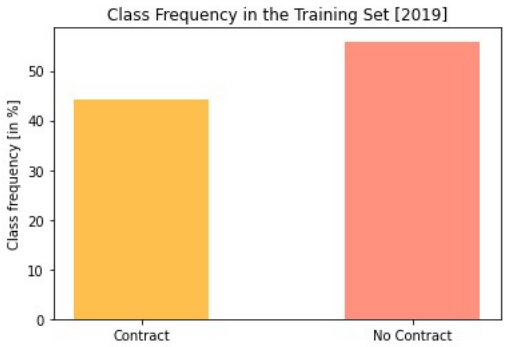

Figure 3.1 Class distribution in 2019 dataset

For our baseline which was the Logistic Regression, we applied the following: Primarily the training and testing set were set up with a test size of 0.30 and a random state of 537. For the pre-processing phase the model used the Standard Scaler and the OneHotEncoder which made it possible to encode all the independent categorical variables into binary columns for each category[1]. Within the pipeline, the Logistic Model was set to have 1000 maximum iterations through the "lbfgs" solver. After fitting our training set and using it to predict our test one, our baseline model calculated the following results:

Evaluation metric	Score
F1 score	0.93
MCC	0.86

Table 3.1 Baseline model scores

Against these results, we followed with the application of the Decision Tree and XGBoost models.

4. Experiments

This section contains the following:

Data: This research employed the dataset deriving from the GEIH - a monthly applied survey in Colombia which aims to provide information related to the labour market, income and monetary poverty, as well as sociodemographic characteristics of the population. The source of the monthly data derives from an open data initiative of the Colombian Government, namely Datos Abiertos.

Out of the information contained in this survey, the final dataset consisted of 39 independent variables and 1 dependent variable, with a total of 504073 observations. The main consideration when designing this dataset was to keep as many columns and observations as possible under the condition of having full and useful information for the prediction[2].

A very interesting phenomenon observed during the data cleaning process, is that the data coming from urban areas was much cleaner and complete than the data from the rural ones, hence making it harder for us as researchers, and conceivably as well for the very government, to be able to extract significant insights from it, such as to identify informal workers in such hard-to-reach zones.

Software: As the research aims to create a fully reproducible machine learning project, open source tools were at the centre of the technical implementation. Google Colaboratory, Anaconda/Jupyter notebooks were used to write and execute arbitrary python code. In addition, tools such as Excel and R were used for data cleaning, processing, etc.

Evaluation method: The main evaluation metrics used during this process were the F1 score, MCC and the confusion matrix. While the F1 score measured the models performance which on average was 0.93, we added an additional layer of evaluation on the performance of our classifications - the MCC, which on our three models calculated an average of 0.83. Further detailed results explained below.

Experimental details: Experiment I - Running models in a smaller sample, first quarter of 2019: Prior to running the models in the final dataset, we worked and ran models on a sample dataset of four months from the final dataset: January, February, March, and April of the year 2019. Python and Jupyter notebooks were used to get the monthly data from the Colombian Government website / GEIH website. From each monthly dataset 6 out of 8 subsets were used, multiplied by 3 geographical divisions of Colombia, where the survey took place.

The subsets are named as follows: State division: Area, Cabecera and Resto. Subsets: Home-housing; General characteristics; Work- force; Unoccupied; Occupied; Inactive; Other activities and unpaid work in the week; Other income. Drop/keep: Out of all subsets, Inactives and Unoccupied were dropped from the experiment.

Rationale: The subsets Inactives and Unoccupied were removed, as, we were interested exclusively in those currently holding a job, as those are the ones over which we are then able to classify as informal or formal workers.

The experimental dataset consisted of around 40,000 observations. Open ended questions, missing values and duplicated columns were dropped together with columns that had more than 50,000 of such invalid values. The y - target variable depended on the question: Given that a person is employed, do they claim to have a contract? - If yes, they were classified as formal workers, and if not they were classified as informal. The performance of the three models that we ran in this sample are shown below:

Evaluation metric	LR	DT	XGBoost
F1 score	0.91	0.96	0.91
MCC	0.82	0.92	0.81

Table 4.1 Performance of models in 2019 quarter dataset

Experiment II - Running models in the complete dataset of 2019: Before moving to the final testing dataset, the three models were also run in a bigger sample, the complete dataset of the year 2019. Based on this dataset the best-performing model was trained and used to predict / classify in the dataset for 2021.

The three aforementioned evaluation metrics were employed to evaluate the models. The following show the results of the Confusion Matrixes for all three models:

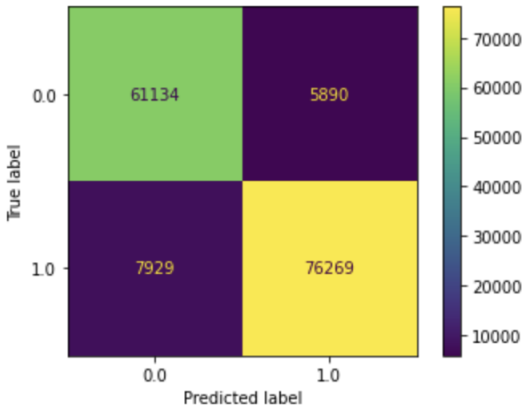

Figure 4.1 Logistic Regression Confusion Matrix - complete 2019 dataset

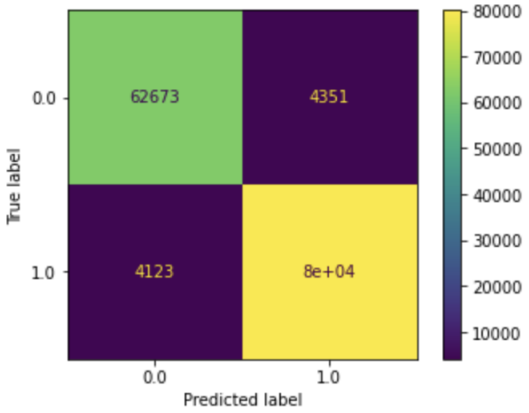

Figure 4.2 Decision Tree Confusion Matrix - complete 2019 dataset

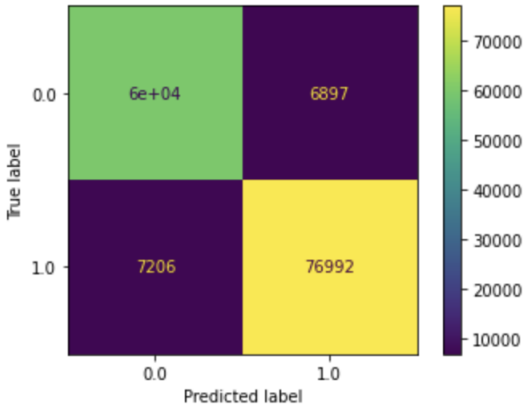

Figure 4.3 XGBoost Confusion Matrix - complete 2019 dataset

Moving along with evaluation metrics, the graph below shows the F1 score and MCC, reafirming the Decision Tree model as the best performing one.

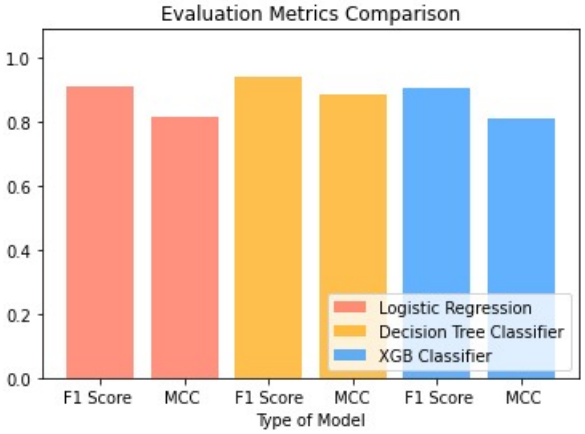

Figure 4.4 Performance of models in 2019 dataset

In the dataset of 2021, the 6 groups that are formed by the characteristics of the working people/respondents and their working environments are separated and evaluated into the the best model, which is the decision tree. According to the evaluation metric F1, the variables that act as a better input for the chosen model are those associated with unexpected expenses and working characteristics. However, when looking at the MCC the model that outperforms all others is the one with the input of working characteristics.

Subset	F1 score	MCC
Unexpected expenses	0.77	0.55
Location	0.56	0.18
Work characteristics	0.89	0.79
Preferences and opinion	0.56	0.17
Sociodemographic characteristics	0.64	0.29
Other activities	0.43	0.10

Table 4.2 Subset categories performace with the DT model

Results: Final phase - Running the best-performing model in the dataset of 2021 and the subsets or categories with predictive power: Prior to applying the predicting model in the testing dataset, the following figure showed the proportion of classes in the 2021 dataset:

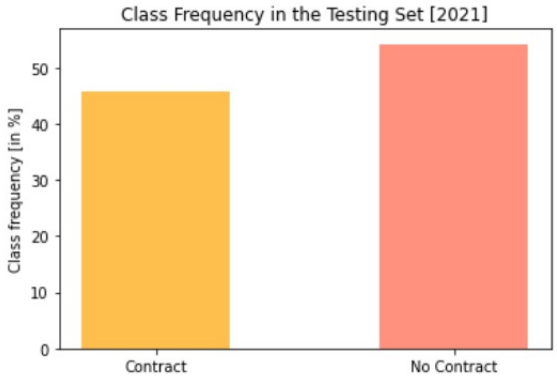

Figure 4.2 Class distribution in 2021 dataset

Commenting the quantitative results: Running the Decision Tree model in the final dataset of 2021, showed the following results:

Evaluation metric	Score
F1 Score	0.89
MCC	0.79

Table 4.3 Performance of DT in the 2021 dataset

The model was run with the 4 variables that make up the working characteristics subset. The evaluation score for this model of the MCC has a score of 0.79. This implies that most of the labels are being predicted correctly.

5. *Analysis

There were several subsets made according to specific characteristics in the working environment, expenses and socio-demographic characteristics. Unexpected expenses account mainly for scenarios of medical expenses, the location subset is related to the person's location, the work location, and if they are in urban or rural areas. Work characteristics, which had the best evaluation metrics is detailed ahead. Preferences and opinion make reference to how satisfied a person is with their work hours and job. Socio-demographic variables are those that describe the person, like sex, relation to the head of the household and education level. Other activities make reference to other activities done in the week.

Unlike how it was initially expected with socio-demographic characteristics having the key role in the prediction, the set of variables with the best scores were those related to the characteristics of the job. These variables are:

- **P6870**: How many employees are there in your workplace?
- **P6880**: In what environment do you mainly do your job (Household, vehicle, office,...)
- **P1881**: What's your main mode of transport for getting to work? (Bus, particular vehicle, bicycle, speedboat)
- **P6240**: What activity did you mainly engage in last week? (Study, work, looking for a job)

Although such variables are not personal on a deep level, they appear to have a big impact when relating them to informal sector workers. As for the first variable, the number of employees ranges from 1 to more than a hundred. This might relate to the capacity of employers to keep formal and informal workers, while it can also relate to entrepreneurship. The environment of a person can conduct their job ranges from households, to vehicles, offices, mines and quarry, the ocean or rivers. In this aspect informal workers tend to have less safe working environments. The main activities the respondents were engaged vary between studying, working and/or looking for a job.

On a macro level, results such as the ones in Table 4.2 and 4.3 show that despite the fact that the model was trained in a dataset of 2019 and tested in the dataset of 2021, between which we saw the manifestation of a pandemic that shaked every economic stability, the model still shows a high predictive power when it comes to our classification question. In addition, on a practical level, discovering what is going on behind the model is of crucial matter, as it opens pathways for alternatives or smaller scale measuring for governments in need of measuring the informal economy with limited resources.

6. *Conclusions

Being context-aware, the GEIH survey represents a very useful tool to understand the composition and particularities of the Colombian informal economy. The main finding of the paper shows that the size of the informal economy is considerably large (more than 50% for both 2019, as well as 2021), which is in line with what would expected based on the known literature, as well as the estimates made by governmental institutions.

Nevertheless, while Colombia's monthly effort to collect information from its citizens and make it available to researchers is to be applauded and encouraged, it is also important to recognize the fact that such initiative is costly, not only in economical terms but also in human labor. This is problematic since the type of countries that could

benefit the most from gaining a deeper understanding of their informal economy tend to be those with less available means to fund such initiatives. Based on the findings discussed in the Analysis section, we come to the conclusion that applying machine learning models in public policy challenges such as the informal economy can support in making the process more cost-efficient and boost the addressing/fighting such phenomenon.

While the limits of the current work, or what could potentially serve the future work in the area is the combination of the subsets to find the combinations with the highest predictive power[10]i, the findings show that further analysis and research in the [which variables/characteristics] can lead to a better understanding of the proportion of the informal economy and the numbers composing it.

In conclusion, further work in the area and embracing the application of machine learning in policy design seems promising in addresing many policy problems that directly and indirectly affect the wellbeing of the citizens, the economy of countries and resultantly their development.

7. Acknowledgements

Being able to access a public and open source dataset has significantly made the research process more efficient as it was not needed to add other layers of communication with the relevant institutions. On a policy level, accessing such data shows an effort toward transparency and accountability but at the same time an incentive for further research in the field. We acknowledge the efforts of the Colombian Government toward open data and encourage further opening of the data, which would allow further research and application of machine learning methods.

8. Contributions

The research period has consisted of a very diverse process that included not only the development of machine learning application skills, but also understanding the contextual aspect of applying such skills in a policy process or in policymaking. As initially foreseen, the tasks related to this paper were divided among three groups, i. Research, both in terms of conducting the literature review, as well as additional sources, ii. Computational Tasks, going from the data collection, to the modeling and building of the final product, and iii. Writing and synthesizing all the technical findings into a compact research paper. On a more specific level the data cleaning and wrangling process was led by Maria Jose Lee with the support of the whole team; the building of the pipelines and models running was led by Carmen Garro with the support of the whole team; and the writing process in synthesizing and interpreting the findings was led by Gresa Smolica with the support of the whole team.

References

[1] Scikit-learn: Machine learning in python, author=Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E., journal=Journal of Machine Learning Research, volume=12, pages=2825–2830, year=2011.

[2] X. Cheng. Preprocessing of categorical predictors in svm, knn and kdc (contributed by xi cheng), Aug 2020.

[3] DANE. Gran encuesta integrada de hogares - geih. https://www.datos.gov.co/Estad-sticas-Nacionales/Gran-Encuesta-Integrada-de-Hogares-GEIH/mcpt-3dws, 2021. Accessed: 2022-10-17.

[4] J. Franic. What do we really know about the drivers of undeclared work? an evaluation of the current state of affairs using machine learning. *AI amp; SOCIETY*, 2022.

[5] C. Hummel. Informal work in numbers. *Why Informal Workers Organize*, page 54–84, 2021.

[6] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013.

[7] C. P. Jorge Alvarez. Covid-19 and the informality-driven recovery: The case of colombia's labor market.

[8] M. W. Loftis and P. B. Mortensen. Collaborating with the machines: A hybrid method for classifying policy documents. *Policy Studies Journal*, 48(1):184–206, 2018.

[9] F. Ohnsorge and S. Yu. The long shadow of informality, Mar 2022.

[10] A. A. Soofi and A. Awan. Classification techniques in machine learning: Applications and issues. *Journal of Basic amp; Applied Sciences*, 13:459–465, 2017.

9. *Appendix

Annex 1. Grouping of variables (to be used for Subset Selection) uploaded in the github repo for better understanding of the subset selection.